EXAMINING THE INFLUENCE OF CORRECT AND INCORRECT "NONE-OF-THE-ABOVE" RESPONSE ALTERNATIVES ON THE POSITIVE AND NEGATIVE TESTING EFFECTS

by

JOSHUA D. KOEN

Presented to the Faculty of the Honors College of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

HONORS BACHELOR OF SCIENCE IN PSYCHOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

November 2007

ACKNOWLEDGMENTS

There are so many people to thank for the guidance and support during this project. First and foremost, I would like to thank my mentor Tim Odegard. His guidance, support, encouragement and trust in me throughout this project has helped me achieve everything I have done. There really are no words to describe how much I appreciate what you have done for me. I wish you the best and look forward to future years of friendship and collaboration. I thank my parents for their love and support throughout my undergraduate college career. They have provided me with so much and helped me through so many rough times in my life. To my grandmother, Ruth "Nan" Cline, who was always there when I needed her. Also, I would like to thank my grandparents, John and Francis Koen, for the love and support they always gave to me. Without the love and support of my family, I would not be the man I am today.

I thank all of my friends for their friendship and support. Darrell Holloway, Josh Sawyer, Allen Medway, Jim Randles, Jason Fisher and others not mentioned have influenced my life. I look forward to years of friendship with all of you. Also, I thank Kara Jenkins, Crystal Cortes, Emily Ferris, John Kretzer, and Serah Obayangban, for their support with this project and the friendship they have provided me. I wish all of you the best and look forward to seeing you in the future. Finally, I would like to thank Denice, whom I love very much, for her love and support during this time.

November 08, 2007

ABSTRACT

EXAMINING THE INFLUENCE OF CORRECT AND INCORRECT
"NONE-OF-THE-ABOVE" RESPONSE ALTERNATIVES
ON THE POSITIVE AND NEGATIVE TESTING EFFECTS

Publication No. _____

Joshua D. Koen, B.S. in Psychology

The University of Texas at Arlington, 2007

Faculty Mentor: Timothy Odegard, Ph.D.

Recently, an abundant amount of research has focused on the benefits and consequences of repeated testing (Roediger & Karpicke, 2006b). Repeated testing has been shown to enhance memory for studied material, a phenomenon called the *positive testing effect*. Another conclusion that has been reached in regards to repeated testing is that there are negative consequences associated with multiple-choice tests, which has been dubbed the *negative testing effect*. However, none of the research focusing on interpolated multiple-choice tests has addressed how inclusive response options, such as "none-of-the-above," influence the positive and negative testing effects.

The present research examines how correct and incorrect "none-of-the-above" response alternatives influence the positive and negative testing effects associated with multiple-choice tests. The present experiments used educationally relevant materials (prose passages; Roediger & Marsh, 2005) to examine this issue. Initially, participants read a set of nonfiction passages. Afterwards, participants completed a multiple-choice test over the passages that they read. Some questions on this test had a "none-of-the-above" response alternative while others did not. The correctness of the "none-of-the-above" response alternative was manipulated so that half of the time this was the correct answer and half of the time it was an incorrect response alternative. After a short filler task, participants received a final cued-recall (Experiment 1) or multiple-choice (Experiment 2) test that contained previously tested questions and questions that were not tested previously (control questions). The results from these two experiments demonstrated that questions with an incorrect "none-of-the-above" response alternative did not have any effect on the positive testing effect. However, the positive testing effect was negated and the negative testing effect was bolstered when questions on the initial multiple-choice test contained a correct "none-of-the-above" alternative.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Many teachers in academic, as well as nonacademic, settings frequently use testing as a tool to assess knowledge learned during the course of instruction. These tests can come in many forms ranging from free recall or essay tests to multiple-choice tests. The goal of education in today's society, such as the *No Child Left Behind* initiative, places a strong emphasis on learning. The view of testing as a means assessing current knowledge is prevalent among many of today's educators, even though a considerable amount of research has demonstrated that tests are not just tools to assess knowledge, but that they actually promote learning and alter memory (e.g., Butler & Roediger, 2007, Glover, 1989; Jones, 1923; Roediger & Marsh, 2005; Spitzer, 1939, for reviews see Roediger & Karpicke, 2006b and Marsh, Roediger, Bjork, and Bjork, 2007). Specifically, testing has been demonstrated to improve the retention of the tested material, a phenomenon known as the *positive testing effect* (e.g., McDaniel & Masson, 1985; Roediger & Karpicke, 2006a), and is by no means new to researchers (e.g., Jones, 1923; Spitzer, 1939). These findings have led many researchers to argue for more frequent testing in classrooms (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Dempster, 1992, 1997; Foos & Fisher, 1988; Glover, 1989). However, most educators have not heeded these suggestions (Dempster & Perkins, 1993)

1.1 The Positive Testing Effect

A study conducted Spitzer (1939) provides a classic example of the positive testing effect. In his study, Spitzer examined the entire sixth grade population of nine Iowa cities. Students in his experiment read a passage and were later tested over it.

Some the students received an immediate test while others received the test at different intervals after the passage was read. The time of the initial test ranged from immediately to 63 days after the passage was read. Additionally, some students who read the passage received multiple retention tests at varying intervals after the initial test. This was done to examine how tests at different intervals would influence subsequent test performance. The results Glover obtained demonstrated that tests given close to the study phase enhanced performance across long delays. Glover concluded that testing enhances retention of previously studied material. The finding that taking a test increases performance on later tests is a robust phenomenon and has been replicated numerous times (e.g., Bruning, 1968; Butler & Roediger, 2007; Carpenter & DeLosh, 2005; 2006; Cull, 2000; Darley & Murdock, 1971; Duchastel & Nungester, 1982; Foos & Fisher, 1988; Glover, 1989; Hogan & Kintsch, 1971; Kang, McDermott, & Roediger, 2007; Karpicke & Roediger, 2007; LaPorte & Voss, 1975; McDaniel & Masson, 1985; McDermott, 2006; Nungester & Duchastel, 1982; Roedgier & Karpicke, 2006a; Rodiger & Marsh, 2005; Tulving, 1967). Additionally, tested material has been demonstrated to have a slower rate of forgetting than material that was only studied (Runquist, 1986a, 1986b; Wheeler, Ewers, and Buonanno, 2003; see however Slamecka & Katsaiti, 1988)

Recently, Karpicke and Roediger (2007) have furthered this notion and demonstrated that frequent testing is important for the long-term retention of material by using a multi-trial learning paradigm. In their first experiment, participants were assigned to different groups that took different amounts of tests during the learning

phase. One group studied material three times and then received a test (SSST group). Another group studied the material then took a test, studied the material again, and took another test (STST group). Finally, a group of participants studied the material once and then took three tests (STTT group). Each group repeated their respective four-phase cycle (SSST, STST, or STTT) five times. Additionally, a final recall test was given to participants one week later. The results demonstrated that testing is not just a tool for assessment because learning occurred across the test phases (see Tulving, 1967). However, performance on the final recall test one week later was greater for participants in the STTT group than in the STST and SSST groups. This finding demonstrates that repeated testing during the learning phase enhances long-term memory. Additionally, Experiment 2 of Karpicke and Roediger's (2007) study demonstrated that this effect is dependent upon whether or not the tests in the learning phase required the recall of all of the material or just the recall of items that had not been recalled yet. They found that tests that only required the recall of previously non-recalled material led to the worst performance on the final retention test. However, performance was best when all of the tests in the learning phase required the recall of all the studied material.

One thing that has not been discussed up to this point is how tests benefit memory in comparison to the common method of studying prescribed by teachers. Commonly, teachers will instruct students to begin studying for an upcoming test one or two weeks before the test is to be given. Students are instructed to study the material for a few hours a day, everyday, until the day of the test. Indeed, research has shown that spaced studying is more beneficial than a single, mass study session (i.e., the *spacing*

effect; Melton, 1970; Underwood, 1970). However, Whitten & Bjork (1977) demonstrated that spacing retrieval (test) attempts in a fashion similar to distributed studying is more beneficial to final memory performance (see also Rea & Modigliani, 1985).

Other research concerning the positive testing effect has focused on the effects of different interpolated tests have on a final retention test. Glover (1989) demonstrated that interpolated free recall tests demonstrate a larger positive testing effect when compared to interpolated cued-recall and multiple-choice tests. Moreover, free recall even showed this benefit regardless of the type of question used on the final retention test. However, it is important to point out that tests such as free recall, cued-recall, short answer, fill-in the blank, and multiple-choice tests have all demonstrated a positive testing effect (e.g., Roediger & Karpicke, 2006a; Carpenteter & DeLosh, 2006; Butler & Roediger, 2007; Foos & Fisher, 1988; Roedger & Marsh, 2005, respectively). However, there is debate about which type of initial tests produces the best benefits to later retention (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Glover, 1989; Foos & Fisher, 1988; Duchastel & Nungester, 1982).

1.2 Theories of the Positive Testing Effect

So far, the findings of the positive testing effect suggest that testing should be used more frequently in classroom settings to promote learning. However, what cognitive processes are responsible for the positive and negative testing effects? Research has attempted to answer this question and has generated two main hypotheses.

First, the additional exposure hypothesis is examined and is followed by the retrieval hypothesis.

1.2.1 Additional Exposure Hypothesis

The additional exposure hypothesis states that testing increases retention by increasing the amount of exposure to the to-be-remembered material (Thompson, Wenger, & Bartling, 1978; Roediger & Karpicke, 2006a). For example, if a person studies a list of words and then is given an immediate recall test, the test allows them an additional opportunity to study the words that they recalled. In contrast, people who do not receive an initial test are only allowed to study the word list once. This hypothesis also predicts that, since testing allows for additional study time, overlearning of the material will occur (Slamecka & Katsaiti, 1988). This hypothesis is supported by findings that show increase in performance on a final test in which participants either studied the material and then completed a filler task (no test condition) or studied the material and took an initial test (test condition; e.g., Roediger & Marsh, 2005).

As discussed by Dempster (1997), the extant literature do not support the additional exposure hypothesis because it suffers from some critical confounds. The major criticism of the hypothesis is that a positive testing effect is still observed even when the amount of exposure to the to-be-remembered material is controlled (e.g., Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a; Tulving, 1967). For example, Roediger & Karpicke (2006a, Experiment 2) had participants learn prose passages. Some participants received alternating study and test phases (STST condition), some received a study phase followed by three consecutive tests (STTT

condition), and others received four study phases (SSSS). Afterwards, all participants received a retention test five minutes, two days, or one week after the learning phase. The results of this experiment demonstrated that participants who received multiple study opportunities performed better on an initial retention test. However, participants who received three tests showed lower rates of forgetting and higher performance on the retention test given two weeks after the initial learning phase. This finding is contradictory to the additional exposure hypothesis because exposure to the to-be-remembered material was controlled for across the three groups. However, exposure to the to-be-remembered material was actually greater in the SSSS condition than the STTT condition, since only recalled material could be rehearsed in the STTT condition.

Nungester and Duchastel (1982) used a similar paradigm to examine the testing effect. In their experiment, participants studied the material twice or studied the material, and then took a multiple-choice test. Their manipulation is important because it actually equated the amount of exposure to the studied material across the two conditions. Their results showed that testing still lead to better performance on a final retention test. Another problem with the additional exposure hypothesis is that it cannot explain why to-be-remembered information, when already stored in memory, is better enhanced by further test trials than study trials (e.g., Brainerd, Kingma, & Howe, 1985).

Additionally, in some instances, testing not only enhances material for the tested information, but can also enhance memory for related, but not tested, material (Chan, McDermott, & Roediger, 2006; Hamaker, 1986). For example, Chan and colleagues (2006) demonstrated that taking a test enhanced recall for initially non-tested material

when the non-tested material was related to the topic being tested. Such findings are problematic for the additional exposure hypothesis because prior testing also enhanced memory for related, yet non-tested, information. These results suggest that something other than additional exposure to the studied material causes the positive testing effect.

1.2.2 Retrieval Hypothesis

The additional exposure hypothesis is only concerned with the amount of exposure to the studied material. As demonstrated above, this hypothesis cannot account for many of the findings regarding the positive testing effect. This hypothesis neglects one important aspect, the act of retrieval. The assumption made by this hypothesis is that retrieval is a passive process with no effects on memory. Instead, the amount and quality of encoding processes determines how long material will be remembered. As discussed by Roediger and Karpicke (2006b), this was a common assumption made by educators and memory researchers alike. Importantly, research has demonstrated that the act of retrieval can modify memory. This had led researchers to propose the retrieval hypothesis of the positive testing effect.

The retrieval hypothesis states that the act of retrieving information from memory causes the positive testing effect (see Roediger & Karpicke, 2006b). In other words, the act of retrieving information for a test will increase accessibility to that same information on subsequent retention tests (e.g., Darly & Murdock, 1971; Nungester & Duchastel, 1982). The act of retrieval is similar to the findings on the generation effect (Slamecka & Graff, 1978). In experiments examining the generation effect, participants are better able to recall words that were generated by them during the initial study phase

than words that were presented to them. Thus, the act of retrieval on an initial test should lead to better performance on the studied material when compared to restudying the same material (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Karpicke & Roediger, 2007; Tulving, 1967; see however Chan & McDermott, in press). The main problem with this hypothesis is that it is broad in regards to the cognitive processes that give rise to the positive testing effect. There are two hypotheses concerned with how initial retrieval leads to the testing effect.

1.2.2.1 Elaborative Retrieval. Carpenter & DeLosh (2006) proposed that the amount of elaboration required on an initial test is paramount to performance on the final test (see also Glover, 1989). In their set of experiments, Carpenter and DeLosh required participants to complete word stems for an initial test. The word stems contained the first letter of the word, the first two letters of the word, the first three letters of the word, or the first four letters of the word. On a final retention test, recall was highest for words that were initially tested with only the first letter. They concluded that the positive testing effect is a function of the amount of elaboration required on the initial test (Gardiner, Craik, & Bleasdale, 1973). As mentioned earlier, Glover (1989) demonstrated that a larger positive testing effect was observed for participants who took an initial free recall test. Taken together, these studies suggest that tests that require more elaborative retrieval demands on an initial test will improve performance on a subsequent retention test (Bjork, 1975; Jacoby, 1978; see however Duchastel & Nungester, 1982). As will be demonstrated in the following experiments, this is not necessarily the case.

1.2.2.2 Accessibility and Retrieval Blocking. Another proposition concerning the influence of retrieval processes on the positive testing effect is that the act of retrieval increases accessibility to the retrieved information (Darley & Murdock, 1971; see Tulving & Pearlstone, 1966, for a discussion of accessibility versus availability). In other words, taking a test over studied material makes it easier for information stored in memory to be accessed on later retrieval attempts. Additionally, this notion can be augmented by the retrieval blocking hypothesis (Raaijmakers & Shiffrin, 1981). This hypothesis is usually implemented in misinformation paradigms; however, it can provide a useful explanation of the testing effect. This hypothesis states that initial retrieval of misinformation blocks access to the accurate information that was stored in memory during the initial study phase (e.g., Schreiber & Sergent, 1998; see however Belli, 1993). With this in mind, it would hold that retrieval correct information on an initial test would block access to erroneous information linked to the test question cue. Taking these two notions together, it can be concluded that the accessibility of items that are tested is increased because, when that item is retrieved, it blocks access to irrelevant, and incorrect, information.

1.2.3 Summary of the Theories

In summary, the extant literature do not support the additional exposure hypothesis. Instead, research has demonstrated that the act of retrieval causes the positive testing effect. However, this hypothesis has not described a specific cognitive process that can account for the effects of prior retrieval. One mechanism by which this can occur is the amount of elaboration afforded to the act of retrieval (Carpenter &

DeLosh, 2006). Additionally, the act of retrieval can block access to competing information and thus increase the accessibility of the correct information (Darley & Murdock, 1971). It is important to note that these two theories of the positive testing effect are not mutually exclusive. The amount of elaboration may increase access to the relevant information by blocking competing information.

1.3 Multiple-Choice Tests

Recently, the effects of multiple-choice tests on the positive testing effect have begun to be more thoroughly investigated (e.g., Butler, Marsh, Goode, & Roediger, 2006; Roediger & Marsh, 2005). Taking an initial multiple-choice test, in most cases, boosts performance on a subsequent test (e.g., Kang, McDermott, and Roediger, 2007; Nungester & Duchastel, 1982; Whitten & Leonard, 1980; see Marsh et al., 2007 for a review). For example, Nungester and Duchastel (1988) demonstrated that taking an initial multiple-choice test lead to better performance on a two-week retention test than studying review statements that contained same information (see however Carpenter & DeLosh, 2006; Glover, 1989).

However, there are negative consequences associated with multiple-choice tests. For example, Roediger and Marsh (2005) demonstrated that participants who received an initial multiple-choice test used incorrect alternatives from this initial test as responses on an immediate final cued-recall test (see also Butler et al., 2006; Whitten & Leonard, 1980). This has been dubbed the *negative testing effect*. However, it is important to point out that forced recall can also lead to a negative testing effect (McDermott, 2006).

Researchers (e.g., Roediger & Marsh, 2005) have discussed that multiple-choice tests are similar to the misinformation paradigm. In a typical misinformation experiment, a person first watches a video or series of slides (e.g., a car running a *yield* sign). Afterwards, the person receives erroneous information embedded within a question that is presupposed to have happened (e.g., How fast was the care going when it ran the *stop* sign?; i.e., misinformation). Finally, participants answer questions about the event (e.g., Did you see a *stop* sign in the film?). Participants are very likely to report the misinformation from the initial question session as having actually been present in the scene (e.g., report that they saw a *stop* sign instead of a *yield* sign; i.e., Loftus, Miller, & Burns, 1978; for a review see Loftus, 2005). This is analogous to common educational practice. First, students are asked to learn some material. In the time course of a class, students are usually tested over a section of material with a multiple-choice test. At the end of a course students usually receive a cumulative final examination. Exposure to incorrect alternatives on an initial multiple-choice examination serves as a source of misinformation.

Indeed, exposure to incorrect information can impair memory for the correct information. For example, being exposed to incorrectly spelled words can lead to more spelling errors (Brown, 1988; Jacoby & Hollingshead, 1990). Previous research has also revealed that false statements on an initial true/false test are rated more truthful at a later point than new false statements (i.e., the *negative suggestion effect*; Remmmers & Remmers, 1926; see also Brown, Schilling, & Hockensmith, 1999; Toppino & Brochin, 1989). Importantly, the negative suggestion effect has also generalized to multiple-

choice test alternatives. Toppino & Luipersbeck (1993) had participants take an initial two alternative multiple-choice test. After a delay, participants in their study were asked to provide truth ratings for multiple-choice alternatives. The incorrect alternative that appeared on the initial multiple-choice test was rated as more truthful than novel incorrect lures. This finding suggests that mere exposure to incorrect alternatives could account for the negative testing effect. However, Roediger and Marsh (2005) demonstrated that this is most likely not the case. The incorrect multiple-choice test alternatives used to answer cued-recall questions in their experiment were more often than not the same incorrect lure selected on the initial multiple-choice test (see also Butler et al., 2006). The research reviewed by Marsh and colleagues (2007) also supports this notion of the negative testing effect. In their review, the authors demonstrated that incorrect multiple-choice alternatives that intruded as responses on a final cued-recall test were more likely due to faulty reasoning processes. In other words, the negative effect of testing is most likely due to a person's erroneous endorsement of an incorrect alternative as the correct response.

1.4 The Present Experiments

Researchers have started to examine the influence of different types of multiple-choice test questions on both the positive and negative testing effects. Roediger and Marsh (2005) demonstrated that questions with more response alternatives decreases the positive testing effect and increases the negative testing effect (see however Butler et al. 2006). Additionally, educators often use inclusive response alternatives, such as "none-of-the-above", on multiple-choice tests. Research in regards to question with a "none-

of-the-above" response alternative is sparse and has mainly focused on difficulty and discrimination ratings (Knowles & Welch, 1992). Additionally, none of the research examining these alternatives has focused on their impact on subsequent test performance. The purpose of the present experiments was to examine initial multiple-choice test questions correct and incorrect "none-of-the-above" response alternatives and their influence on both the positive and negative testing effect.

Researchers examining eyewitness identification memory have argued that commitment to an initial mug shot impairs memory for the actual perpetrator of the crime (Dysart, Lindsay, Hammond, & Dupuis, 2001; Gorenstein & Ellsworth, 1981, see Deffenbacher, Bornstein, & Penrod, 2006, for a meta-analytic review). For example, Gorenstein & Ellsworth (1980) found that selecting an incorrect suspect from an initial photo array impaired memory for the actual perpetrator on a final identification task. Additionally Schooler, Foster, and Loftus (1988) showed that tests that promote incorrect responding are detrimental to subsequent test performance. In their experiments, participants studied a slide sequence that depicted a scene of a wallet being snatched. Afterwards, their participants received an initial two-alternative forced-choice recognition test. Some participants received a recognition test with the correct alternative and an incorrect alternative, others received a test with two incorrect alternatives, and others did not receive the initial test. Participants were given a final retention test immediately after the filler task. Participants who were forced to make an incorrect response were less likely to favor the correct answer on the final retention test than participants who did not receive an initial multiple-choice test (Schooler et al.

1988, Experiment 1). Furthermore, these researchers demonstrated that this effect held even when the initial incorrect lures from the first test were not present on the final retention test (Schooler et al., 1988, Experiment 2).

In both of the studies described above, the correct answer was omitted from the initial test. As discussed by Schooler and colleagues (1988), these types of tests are analogous to test that include a correct "none-of-the-above" alternative. They argued that these types of tests promote incorrect responding and would be detrimental to subsequent test performance (see also Gross, 1994). However, these studies are limited in regards to the present research question since they did not include a "none-of-the-above" alternative. Participants in their experiments were forced to select an incorrect answer.

As noted by Knowles and Welch (1992), research regarding the use "none-of-the-above" as a response alternative has yielded mixed results. Their meta-analysis demonstrated that multiple-choice questions that have a "none-of-the-above" alternative are similar in difficulty to questions that do not contain this response alternative. As discussed by Wells (1993), witnesses to experimentally created crimes are more likely to select a suspect in a line-up even when the actual perpetrator is not included in the line-up and the witnesses are told that the perpetrator may or may not be in the line-up (see Clark & Davey, 2005). These findings suggest that participants will be more likely to select an incorrect lure on a multiple-choice test when the question contains "none-of-the-above" as the most appropriate response alternative.

The findings discussed above also provide a further explanation for the positive testing effect. The findings from Schooler and colleagues (1998; Experiment 1) suggest that the correct answer must be present on an initial multiple-choice test to observe a positive testing effect on a final retention test. Indeed, much of the research using multiple-choice tests to examine the positive testing effect have used multiple-choice questions that contain the correct answer (e.g., Roedigner & Marsh, 2005; Whitten & Leonard, 1980). However, none of the research reviewed up to this point has directly examined the positive testing effect by using multiple-choice questions that do not contain the correct answer.

The paradigm used by Roediger and Marsh (2005) was adapted for use to examine if including correct and incorrect "none-of-the-above" response alternatives would influence the testing effect. Participants in both of the experiments reported herein studied a set of non-fiction passages and then received an initial multiple-choice test. The initial multiple-choice test included three different types of questions. One of the question types contained the correct answer and three incorrect lures (A-D questions). An incorrect "none-of-the-above" response alternative was added to A-D questions to create the second question type (A-E "E" Incorrect questions). The third type of question contained four incorrect alternatives and a correct "none-of-the-above" response alternative (A-E "E" Correct questions). Examples of these questions can be found in Table 2.1. After the initial multiple-choice test, participants completed a short filler task and then took a final cued-recall (Experiment 1) or multiple-choice

(Experiment 2) test. This final test contained the same questions that were tested on the initial test as well as questions that were not tested (control questions).

1.4.1 Hypotheses

There were three main hypotheses formulated based on the research reviewed above. First, it is predicted that initial multiple-choice questions that contain the correct answer (A-D and A-E "E" Incorrect) will demonstrate a positive testing effect. Specifically, correct performance on the final cued-recall and multiple-choice test will be higher for those questions that were initially tested with questions that contained the factually correct answer as a response alternative. This prediction is based on the findings that multiple-choice tests that contain the correct answer improve performance on a final retention test (e.g., Nungester & Duchastel, 1982; Roediger & Marsh, 2005; Whitten & Leonard, 1980). Moreover, it is predicted that initial multiple-choice questions with "none-of-the-above" as the most appropriate response alternative will not demonstrate a positive testing effect. In other words, correct performance on the final cued-recall and multiple-choice test questions initially tested with a correct "none-of-the-above" response alternative should be similar, or significantly worse, than final retention test questions that were not initially tested. This is predicted because participants will most likely prefer to endorse an informational response alternative (an alternative that contains a plausible answer choice that may accurately complete the question) than reject all of the plausible alternatives (see Wells, 1993). Additionally, this type of responding should lead to impaired performance on the final retention test (e.g., Schooler et al., 1988). This finding would support the retrieval blocking

hypothesis. However, it may be the case that a larger testing effect is observed for questions initially tested with a correct "none-of-the-above" response alternative. This finding would support the elaborative retrieval hypothesis of the testing effect since more elaboration is required to generate the correct answer on these types of questions than on questions where the correct answer is a response alternative.

In regards to the negative testing effect, I predicted that questions with a correct "none-of-the-above" response alternative would increase the amount of lure intrusions on the final retention test, thus, increasing the negative testing effect. This was predicted based on the review provided by Marsh and colleagues (2007) and the accessibility and retrieval blocking hypothesis. Participants will be more likely to provide incorrect alternatives as responses to final test questions because they will be more likely to select an incorrect answer on initial multiple-choice questions with "none-of-the-above" as a correct alternative. Furthermore, this is predicted by the retrieval blocking hypothesis because initial retrieval of erroneous information has been demonstrated to block access to the correct information. Additionally, this is augmented by evidence demonstrating that students are unlikely to change their answer even when they are given an opportunity to do so (Higham and Gerrard, 2005).

2. EXPERIMENT 1

2.1 Methods

2.1.1 Participants

Thirty-two undergraduate students from the University of Texas at Arlington participated in this experiment for partial fulfillment of a course requirement. Participants were run individually or in groups of two. Consent was obtained from each participant.

2.1.2 Design

The design of this experiment conformed to a 2 (passage status: read, unread) X 2 (test type: A-D, A-E) X 2 (item type: critical, control) mixed factorial design. Additionally, a 2 (A-E question type: correct, incorrect) within participant variable was embedded in the A-E level of the test type variable.

2.1.3 Materials

The same 36 nonfiction passages from the Graduate Record Examination (GRE) and the Test of English as a Foreign Language (TOEFL) practice test books used by Roediger and Marsh (2005) were adapted for use in the present experiment. The 36 nonfiction passages were divided into two groups of 18 passages. Only one set of passages was given to participants to read during the study phase. The two sets of passages served as read and unread passages an equal number of times across participants.

The multiple-choice and cued-recall tests were adapted for use in the present experiment from the materials created by Roediger and Marsh (2005). From these

materials, there were a total of 144 multiple-choice questions with six response alternatives. One of the response alternatives was the correct answer while the other five were incorrect lures. Three of the incorrect lures from the five provided by Roediger & Marsh were randomly selected for use in this experiment. A paper-and-pencil multiple-choice test was created. A total of 72 questions out of the 144 total questions appeared on the initial multiple-choice test. Of the 72 multiple-choice test questions, 36 were concerned with information from read nonfiction passages (read questions) and the other 36 covered information from unread nonfiction passages (unread questions).

An A-D and A-E multiple-choice test format was created to test the hypotheses. An example of each question type appears in Table 2.1. For the A-D multiple-choice test format, each question (e.g., *What was Louis Armstrong's nickname?)* had four response alternatives. One of them was the correct answer (i.e., *Satchmo*) while the other three served as incorrect lures (e.g., *Babs, Bird, and King*). A fifth, "none-of-the-above" response alternative was added to every question to create the A-E multiple-choice test format. Furthermore, half of the 36 read and unread questions on the A-E test format had the correct answer present (i.e., *Satchmo*; A-E "E" Incorrect questions) and the other half had "none-of-the-above" as the correct response alternative (A-E "E" Correct questions). An additional incorrect lure was taken from Roedger & Marsh's (2005) to create the A-E "E" Correct question type. This extra incorrect lure replaced the correct answer (e.g., *Satchmo* was replaced by *Fletcher*) to make "none-of-the-above" the correct resposne.

An additional 22 filler questions were added to each multiple-choice test so that each test would contain 94 questions. Two forms of the A-D test condition were created for counterbalancing purposes. Four forms of the A-E test were created so that each question would serve as a critical and control question, as well as an "E" Incorrect and "E" Correct question, an equal number of times across participants.

Table 2.1. *Examples of the three question types for the question "What was Louis Armstrong's Nickname?"*

	Type of Question		
Response Alternative	A-D	A-E "E" Incorrect	A-E "E" Correct
A	Babs	Babs	Babs
B	**Satchmo**	**Satchmo**	Fletcher
C	Bird	Bird	Bird
D	King	King	King
E	-	None-of-the-above	**None-of-the-above**

Note. The correct answers to the question are in bold.

A paper-and pencil cued-recall test was also created from the Roediger & Marsh (2005) stimuli. The cued-recall questions were the exact same as the multiple-choice questions with the exception that they had no response alternatives (e.g. *What was Louis Armstrong's nickname?* _____). The cued-recall test contained 216 questions. Of these questions, 72 were from read passages, 72 were from unread passages, and 72 were

filler questions. For participants in the A-D test condition, half of the read and unread questions were tested on the initial multiple-choice test (critical questions) and half had not been tested on the initial multiple-choice test (control questions). This was the same for the A-E test condition with a slight modification. Half of the 36 read and 36 unread critical questions were initially tested with the "E" Incorrect format while the other half were initially test in the "E" Correct format. The 72 filler questions contained the same 22 filler questions from the initial multiple-choice test. However, none of the filler questions had any corresponding passages in the materials and were not included in the data analysis.

2.1.3 Procedure

The procedure used in the present experiment was similar to the one used by Roediger & Marsh (2005). There were four phases in this experiment. The first phase consisted of reading a set of 18 passages. During this phase, participants were given a total of 90 s to read each passage. Additionally, participants were given a reading log and asked to place a check mark next to the passage number when they completed reading it. This was done to ensure that participants read each passage only once. The second phase was the multiple-choice test phase. Participants were randomly assigned to one of the two test type conditions (A-D or A-E). Participants were instructed that they would have 25 min to complete the test. Participants were further instructed to go through each question in order without going back to a question and to guess if they did not know the answer.

Participants received a trail-making task (Armitage, 1945) that lasted for 5 min and served as the filler task. This task required participants to connect the dots according to a rule (i.e., *1* to *2*, *2* to *3*; *1* to *A*, *A* to *2*, *2* to *B*). Participants were told that they were to connect the dots as fast as possible because they would be timed.

After the filer task, participants received the cued-recall test. Participants were instructed that they would have 35 min to complete this test. Unlike the initial multiple-choice test, participants received a strong warning against guessing. Participants were instructed to place an "X" in the blank if they did not know the answer. After participants completed the cued-recall test, they were debriefed and thanked for their participation.

2.2 Results and Discussion

There were two main hypotheses for this experiment concerning the addition of a "none-of-the-above" alternative on an initial multiple-choice test. First, it was predicted that questions initially tested with a correct "none-of-the-above" response alternative would negate the positive testing effect. However, when "none-of-the-above" was an incorrect response, a positive testing effect should be evident. To examine this, the proportion of correct cued-recall responses was analyzed (see Table 2.2). Furthermore, questions with a correct "none-of-the-above" alternative on the initial multiple-choice test should increase the negative testing effect. The means and standard deviations for the proportion of incorrect lures that were imported from the initial multiple-choice test onto the final cued-recall test can be found in Table 2.3. Finally, performance on the initial multiple-choice test was examined to determine if

there were differences in performance between the three types of initial test questions. All analyses were significant at an alpha level equal to or less than .05 unless otherwise specified.

Table 2.2. *Proportion of correct responses on the final cued-recall test as a function of passage status, item type, and multiple-choice question type.*

	Passage Status	
Item type X MC Question Type	Read	Unread
Critical		
A-D	.46 (.21)	.19 (.13)
A-E "E" Incorrect	.38 (.19)	.19 (.10)
A-E "E" Correct	.24 (.17)	.06 (.07)
Control		
A-D	.30 (.16)	.09 (.07)
A-E	.28 (.14)	.08 (.05)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

2.2.1 Correct Cued-Recall Performance.

The proportion of correct cued-recall responses can be found in Table 2.2. The main hypothesis regarding the positive testing effect would be supported if performance on cued-recall questions initially tested as A-E "E" Correct questions did not significantly differ or was significantly lower than control questions. Additionally,

questions that contain the correct answer (A-D and A-E "E" Incorrect) on the initial multiple-choice tests should be answered with significantly more correct responses than control questions. Correct performance on two types of A-E questions were each compared to performance on A-D questions due to the complexity of this incomplete design.

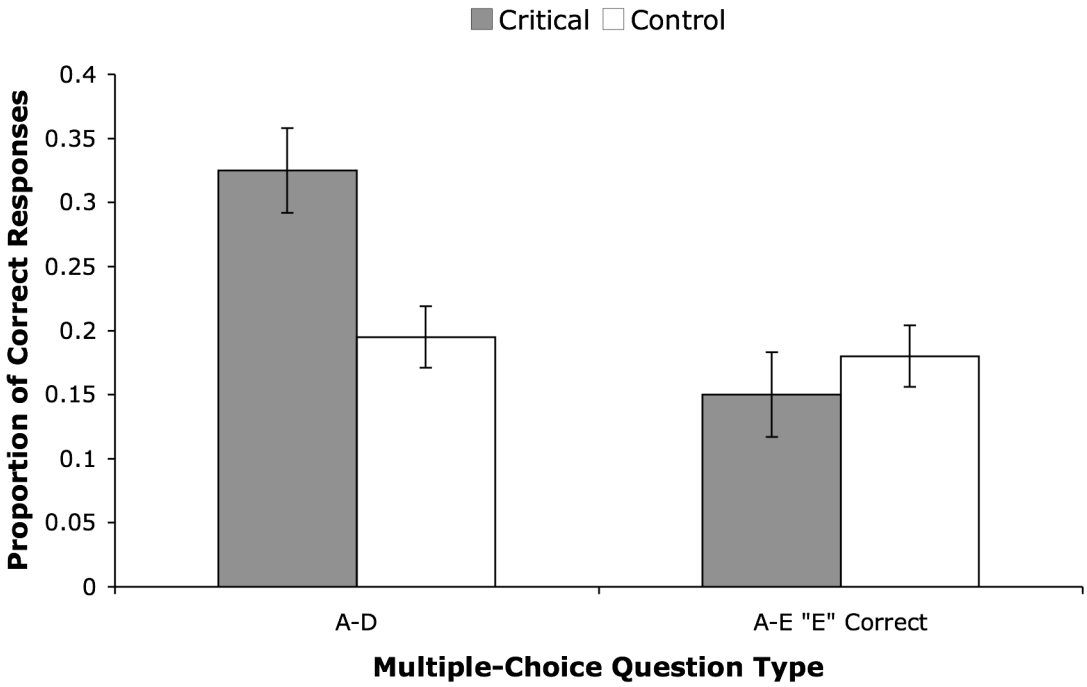

Figure 2.1. *The marginal means (with standard error bars) for the interaction between item-type and multiple-choice question type for the ANOVA comparing A-D and A-E "E" Correct questions.*

An initial 2 (passage status: read, unread) X 2 (item type: critical, control) X 2 (multiple-choice question type: A-D, A-E "E" Incorrect) mixed factorial ANOVA was

conducted to examine whether or not the addition of an incorrect "none-of-the-above" response alternative had any impact on the testing effect. As would be expected, there was a main effect of passage demonstrating that questions taken from read passages (*M* = .36, *SD* = .18) were answered with more correct responses than questions taken from unread passages (*M* = .14, *SD* = .09), $F(1,30) = 79.28$, *MSE* = .019. Also, participants performed better on questions that were tested on the initial multiple-choice test (critical questions: *M* = .31, *SD* = .16) than on questions that were not tested previously (control: *M* = .19, *SD* = .11), $F(1,30) = 63.89$, *MSE* = .007. There was no significant effect of multiple-choice question type nor any interactions, all *F*'s < 1.28. These results demonstrate that the addition of incorrect "none-of-the-above" response alternatives does not have any impact on the positive testing effect.

An additional 2 (passage status: read, unread) X 2 (item type: critical, control) X 2 (multiple-choice question type: A-D, A-E "E" Correct) mixed factorial ANOVA was conducted to examine if removing the correct answer from the multiple-choice test questions negatively influenced the positive testing effect. This ANOVA revealed that participants correctly answered more questions that were tested previously (critical questions: *M* = .24, *SD* = .15) than questions that were not (control questions: *M* = .19, *SD* = .11), $F(1,30) = 16.14$, *MSE* = .006. There was also a main effect of multiple-choice question type, $F(1,30) = 6.05$, *MSE* = .049, that was qualified by a significant interaction between multiple-choice question type and item type, $F(1,30) = 36.52$, *MSE* = .006. This interaction is depicted in Figure 2.1. Participants who received the A-D test format performed better on questions that were tested on the initial multiple-choice test

(*M* = .33, *SD* = .17) than on questions that were not tested on the initial multiple-choice test (*M* = .20, *SD* = .12), *F*(1,30) = 50.61, *MSE* = .006. However, participants performed statistically similar on questions initially test as A-E "E" Correct questions (*M* = .15, *SD* = .12) and control questions (*M* = .18, *SD* = .10), *F*(1,30) = 2.05, *MSE* =.006, $p > .05$. This result supports the main hypothesis by demonstrating that initial multiple-choice test questions with a correct "none-of-the-above" response alternative impaired performance on the final cued-recall test.

Table 2.3. *Mean proportion of incorrect lures that were imported onto the cued-recall test as a function of passage status, item type, and multiple-choice test question type.*

	Passage Status	
Item Type X MC Question Type	Read	Unread
Critical Questions		
A-D	.10 (.07)	.10 (.09)
A-E "E" Incorrect	.15 (.09)	.08 (.09)
A-E "E" Correct	.16 (.12)	.15 (.11)
Control Questions		
A-D	.04 (.04)	.02 (.03)
A-E	.04 (.04)	.04 (.03)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

2.2.2 Incorrect Lures used on the Cued-Recall Test.

The means and standard deviations for the proportion of incorrect lures imported onto the cued-recall test can be found in Table 2.3. If the hypothesis regarding the negative testing effect is supported, then participants will use significantly more incorrect lures from the initial multiple-choice test as answers to final cued-recall questions initially tested with a correct "none-of-the-above" alternative than questions that were not tested on the initial multiple-choice test.

Initially, a 2 (passage status: read, unread) X 2 (multiple-choice question type: A-D, A-E "E" Incorrect) X 2 (item type: critical, control) mixed factorial ANOVA was conducted on the proportion of incorrect lures from the initial multiple-choice test that were used as responses to final cued-recall questions. As expected, participants answered with more incorrect lures on critical (M = .11, SD = .09) than on control (M = .04, SD = .04) questions, $F(1,30) = 26.67$, MSE = .006. The effect of multiple-choice question type was not significant, $F < 1$. However, there was a significant two-way interaction between passage status and multiple-choice question type, $F(1,30) = 6.10$, MSE = .001. Participants used more incorrect lures as answers on read cued-recall questions initially tested in the A-E "E" Incorrect format (M = .10, SD = .07) than on questions initially tested in the A-D format (M = .06, SD = .06), $F(1,30) = 16.88$, MSE = .001. There was no difference between the two question types in the rate of incorrect lures used on unread cued-recall questions, $F < 1$. This interaction was also qualified by a significant three-way interaction, $F(1,30) = 7.60$, MSE = .002. Participants used more incorrect lures as responses to questions initially test as A-E "E" Incorrect questions

that were derived from read passages (M = .15, SD = .09) as opposed to unread passages (M = .08, SD = .09), $F(1,30) = 24.50$, MSE = .002. However, this was not the case for A-D critical questions and control question, both F's < 1.

An additional 2 (passage status: read, unread) X 2 (multiple-choice question type: A-D, A-E "E" Correct) X 2 (item type: critical, control) mixed factorial ANOVA was conducted to assess how a correct "none-of-the-above" response alternative influenced the use of incorrect lures as responses to cued-recall questions compared to the A-D test condition. There was no effect of passage status or any interaction with passage status, all F's < 1. There was a main effect of multiple-choice question type, $F(1,30) = 4.66$, MSE = .008. Participants used more incorrect lures as answers to questions that had "none-of-the-above" as a correct response alternative (M = .10, SD = .08) when compared to the A-D condition (M = .07, SD = .06). The two-way interaction between item type and multiple-choice question type was not significant, $F(1,30) = 2.45$, MSE = .006, p = .13. Finally, a 2 (passage status: read, unread) X 3 (multiple-choice question type: A-E "E" Incorrect, A-E "E" Correct, control) repeated measures ANOVA was conducted to see if the negative testing effect was differentially influenced by the two types of "none-of-the-above" questions. This ANOVA revealed a main effect of multiple-choice question type, $F(1,15) = 16.44$, MSE = .006. Simple contrasts revealed that participants used more incorrect lures on A-E "E" Correct (M = .12, SD = .09) and A-E "E" Incorrect (M = .16, SD = .12) questions than on control (M = .04, SD = .04) questions, $F(1,15) = 11.37$, MSE = .013; $F(1,15) = 25.52$, MSE = .015, respectively. Also, participants used more incorrect lures as responses to cued-recall

questions that were initially tested in the A-E "E" Correct format than on questions that were initially test in the A-E "E" Incorrect format, $F(1,15) = 7.33$, $MSE = .008$.

2.2.3 Initial Multiple-Choice Test Performance

Performance on the initial multiple-choice was a secondary question addressed in this experiment. However, these results demonstrate how well participants performed on the different types of questions on the initial test. The mean proportion of correct responses provided by participants on the initial test is presented in Table 2.4.

Table 2.4. *Proportion of correct responses on the initial multiple-choice test in Experiment 1.*

	Passage Status	
MC Question Type	Read	Unread
A-D	.69 (.15)	.41 (.09)
A-E "E" Incorrect	.55 (.16)	.38 (.09)
A-E "E" Correct	.36 (.18)	.21 (.14)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

First, correct performance between the A-D and A-E "E" Incorrect questions was compared by conducting a 2 (passage status: read, unread) X 2 (multiple-choice question type: A-D, A-E "E" Incorrect) mixed factorial ANOVA on the proportion of correct responses on the initial multiple-choice test. There was a significant interaction between the two independent variables, $F(1,30) = 4.47$, $MSE = .01$ (see Table 4).

Performance on questions derived from read passages was lower when "none-of-the-above" was an incorrect response alternative than when it was not a response alternative, $F(1,30) = 15.80$, $MSE = .01$. However, performance on unread questions did not differ between the A-D and A-E "E" Incorrect conditions, $F(1,30) = 1.00$, $MSE = .01$, $p > .05$.

Another 2 (passage status: read, unread) X 2 (multiple-choice question type: A-D, A-E "E" Correct) mixed factorial ANOVA was also performed on the proportion of correct multiple-choice test responses to compare performance between the A-D test and A-E "E" Correct questions on the A-E test. There was also a significant interaction between passage status and test type, $F(1,30) = 8.32$, $MSE = .009$. Post hoc analyses confirmed that performance was lower on A-E "E" Correct questions derived from read passages ($M = .36$, $SD = .18$) than on A-D questions ($M = .69$, $SD = .15$), $F(1,30) = 31.55$, $MSE = .028$. Additionally, correct performance on A-E "E" correct questions taken from unread passages ($M = .21$, $SD = .14$) was lower than performance on A-D questions ($M = .41$, $SD = .09$), $F(1,30) = 22.70$, $MSE = .014$. This effect was larger for questions derived from read passages. Finally, a 2 (passage status: read, unread) X 2 (multiple-choice question type: A-E "E" Incorrect, A-E "E" Incorrect) repeated measures ANOVA. This ANOVA revealed a main effect of test type, $F(1,15) = 46.06$, $MSE = .011$. Participants performed better on questions when the factually correct answer to the question was present ("E" Incorrect: $M = .47$, $SD = .13$) than on questions where "none-of-the-above" was the correct response ("E" Correct: $M = .29$, $SD = .16$).

2.2.4 Summary of Experiment 1

There were three main findings in the present experiment. First, the positive testing effect was negated when questions were initially tested with a correct "none-of-the-above" alternative, supporting the hypothesis. Additionally, these same questions demonstrated a larger negative testing effect. Participants answered cued-recall questions with more incorrect lures from A-E "E" Correct questions on the initial multiple-choice test. Additionally, performance on initial multiple-choice questions with a correct "none-of-the-above" alternative was lower than questions with incorrect "none-of-the-above" alternatives.

It is important to point out a few limitations in the design of this experiment. First, one test format did not have any "none-of-the-above" response alternatives (A-D test) while the other (A-E test) had a "none-of-the-above" response alternative on every question. This manipulation could have led participants to not give "none-of-the-above" responses on the initial multiple-choice test because it was not a salient feature of the test. Additionally, the final test did not require participants to give a response. This manipulation could have led to dramatically different results on cued-recall performance if participants were forced to respond (for a discussion of this see Higham, 2002). Finally, the use of a final cued-recall test is not very ecologically valid since most educators stay with a similar test format throughout a course. These issues are addressed in Experiment 2.

3. EXPERIMENT 2

The purpose of this experiment was to replicate and extend the results obtained in Experiment 1 to a final multiple-choice test. An additional aim of this experiment was to replicate the findings of Schooler et al. (1988) and extend them to the present paradigm. Schooler and colleagues demonstrated that when participants were forced to give an incorrect response, performance on the final retention test was impaired. Additionally, Experiment 2 of their study demonstrated that this impairment is still present when the initial incorrect alternatives were replaced with the correct answer and a new, incorrect lure. To do so, a multiple-choice test was used as the final retention test. All questions on this final multiple-choice test contained the correct answer and three incorrect alternatives. Additionally, some participants received a final test with the same incorrect lures as the initial multiple-choice test (same lure type) and some received the final multiple-choice test with different incorrect lures than the initial test (different lure type). Based on the findings of Schooler and colleagues (1988), as well as the results obtained from Experiment 1, it is predicted that questions initially tested with a correct "none-of-the-above" alternative will not show a positive testing effect on the final multiple-choice test. This should occur even when the incorrect lures on the final multiple-choice test are different from the initial multiple-choice test.

3.1 Method

3.1.1 Participants

Thirty-two undergraduates from the University of Texas at Arlington participated in this experiment for partial fulfillment of a course requirement.

Participants were tested in groups of no more than two. Informed consent was obtained from all participants.

3.1.2 Design

The design of this experiment conformed to a 2 (passage status: read, unread) X 4 (question type: A-D, A-E "E" Incorrect, A-E "E" Correct, Control) X 2 (lure type: same, different) mixed factorial design with the first two variables manipulated within-participants and the latter manipulated between participants.

3.1.3 Materials

The materials used in this experiment were similar to Experiment 1 with some differences within the two tests. All of the six response alternatives to each question that were generated by Roediger & Marsh (2005) were adapted for use in this experiment. Two additional incorrect alternatives were generated and thus, created a total of one correct answer and seven incorrect lures for the response alternatives for each multiple-choice test question. Three of the seven incorrect lures were randomly chosen and served as incorrect lures to initial multiple-choice test questions for all participants. Furthermore, a fourth incorrect lure was randomly selected from the remaining four incorrect alternatives and was substituted for the correct answer on A-E "E" correct questions.

The initial multiple-choice test contained a total 130 questions. Fifty-four of these questions were from read passage and 54 were from unread passages. Each set of 54 questions was divided into three groups of 18 questions. One set of 18 questions was presented in the A-D question format, another set of 18 questions was presented in the

A-E “E” Incorrect format, and the remaining set of 18 questions appeared in the A-E “E” Correct format (see Table 1 for an example of each question type). An additional 22 questions with no corresponding passages served as filler questions on the initial multiple-choice test. There were 36 questions (18 read and 18 unread) that were not tested on this test and served as control questions. Four different test orders were created so that each questions served in each of the question types an equal number of times across participants.

The final test was changed from a cued-recall test to a multiple-choice test. This test contained 216 questions that consisted of 72 questions derived from read passages, 72 questions derived from unread passages, and 72 filler questions. Within each of the 72 read and 72 unread questions, 18 were tested in the A-D format on the initial multiple-choice test, 18 were initially tested in the A-E “E” Incorrect format, 18 were initially test in the A-E “E” Correct format, and the remaining 18 were not tested on the initial multiple-choice test and served as a baseline measure of performance. As in Experiment 1, the 72 filler questions had no corresponding passages and were excluded from the data analysis. The questions on the final multiple-choice test all contained the correct answer and three incorrect lures. The incorrect lures were either the same as those used on the first test (same lure type) or replaced with the three additional lures that were not presented on the first test (different lure type).

3.1.4 Procedure

The procedure was the same as in Experiment 1 except that participants were instructed to guess on the final multiple-choice test.

3.2 Results and Discussion

The primary question of this experiment is to examine if the detrimental effect of "none-of-the-above" response alternatives obtained with a final cued-recall test in Experiment 1 would generalize to a final multiple-choice test. If this is the case, then correct performance on final multiple-choice test questions that were initially tested with a correct "none-of-the-above" alternative would be similar to or lower than questions that were not tested on the initial multiple-choice test. Additionally, another purpose of this experiment was to examine if the inclusion of new, incorrect lures on the final multiple-choice test would attenuate the detrimental effect of correct "none-of-the-above" alternatives on the positive testing effect observed in Experiment 1. Thus, correct performance on the final multiple-choice test is examined first. If the results from Experiment 1, as well as the results obtained by Schooler, Foster, and Loftus (1988; Experiment 2),are replicated, then even changing the incorrect lures form the initial multiple-choice test to the final multiple-choice test should still show no evidence of a positive testing effect.

Additionally, the negative testing effect was again examined. This was not examined in the same was as in Experiment 1 and is discussed before the results on this dependent variable are present. Finally, initial multiple-choice test performance was examined to see if the results from Experiment 1 would be replicated.

3.2.1 Final Multiple-Choice Test Performance

The mean proportion (and standard deviations) of correct responses given to final multiple-choice questions can be found in Table 3.1. The hypothesis regarding the

positive testing effect and correct "none-of-the-above" alternatives would be supported if performance on final multiple-choice test questions that were initially test with a correct "none-of-the-above" alternative is similar to performance on final multiple-choice questions that were not initially tested. Additionally, this trend in performance should not be moderated by lure type. Specifically, performance on A-E "E" Correct questions should not be significantly higher than performance on control questions for both the same and different lure type conditions.

Table 3.1. *Proportion of correct responses on the final multiple-choice test in Experiment 2 as a function of passage status, lure type, and multiple-choice question type.*

	Passage Status X Lure Type			
	Same Lures		Different Lures	
MC Question Type	Read	Unread	Read	Unread
A-D	.68 (.19)	.43 (.15)	.77 (.13)	.51 (.14)
A-E "E" Incorrect	.74 (.13)	.44 (.12)	.74 (.16)	.54 (.12)
A-E "E" Correct	.50 (.17)	.34 (.15)	.64 (.20)	.39 (.15)
Control	.64 (.19)	.40 (.16)	.65 (.20)	.40 (.12)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

To examine this a 2 (passage status: read, unread) X 4 (initial multiple-choice question type: A-D, A-E "E" Incorrect, A-E "E" Correct, control) X 2 (lure type: same, different) mixed factorial ANOVA was conducted on the proportion of correct responses on the final multiple-choice test. As expected, participants performed better on questions taken from read passages (M = .67, SD = .17) than questions taken from unread passages (M = .43, SD = .14), $F(1,30) = 137.52$, MSE = .027. Of most importance, the ANOVA yielded a significant main effect of question type, $F(3, 90) = 17.39$, MSE = .017. This main effect is visually depicted in Figure 3.1. Simple contrasts confirmed that questions initially tested in the A-D format (M = .60, SD = .15) were answered with more correct answers than questions not initially tested (M = .52, SD = .17), $F(1,30) = 13.65$, MSE = .027. This result demonstrates a positive testing effect. Furthermore, questions initially tested in the A-E "E" Incorrect format (M = .62, SD = .13) also demonstrated a positive testing effect, $F(1,30) = 18.23$, MSE = .03. However, participants performed significantly worse on A-E "E" Correct questions (M = .47, SD = .17) than control questions, $F(1,30) = 4.44$, MSE = .045. This result replicates and extends the results obtained in Experiment 1 by showing that questions with a correct "none-of-the-above" alternative on an initial multiple-choice test negates the positive testing effect. Importantly, there was no interaction between initial multiple-choice question type and lure type, $F < 1$.

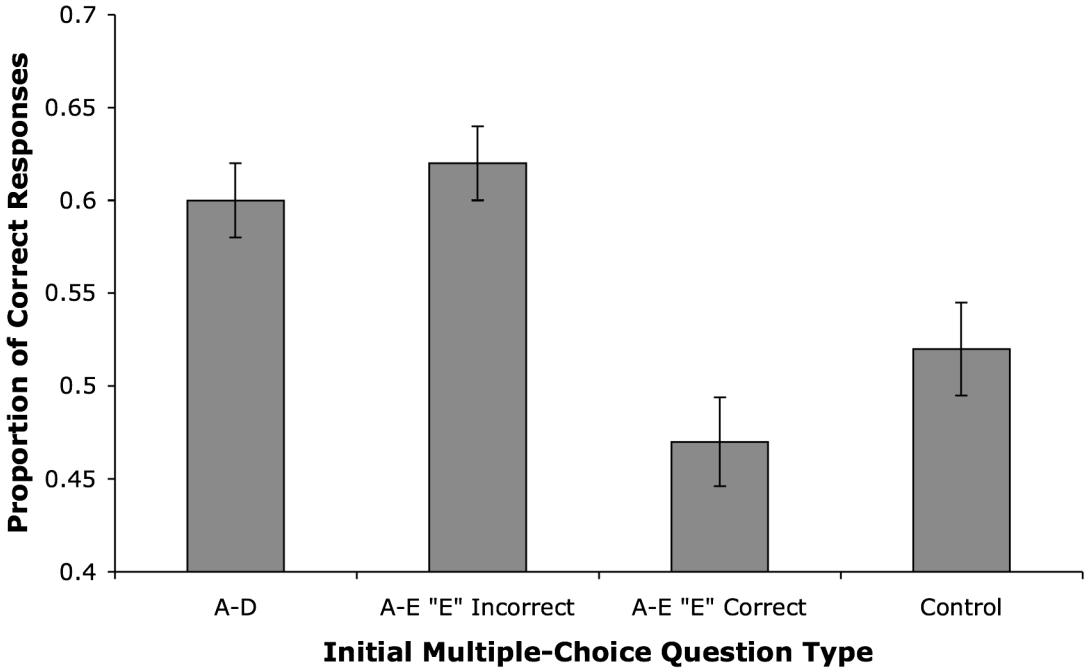

Figure 3.1. *Mean proportion (with standard error bars) of correct performance on the final multiple-choice test collapsed across passage status and lure type.*

3.2.2 Same Lure Selection on both multiple-choice tests

As stated earlier, the negative testing effect was examined using a different measure from Experiment 1. In this experiment, the proportion of final multiple-choice test questions that received an incorrect response that was the same as the incorrect response on the initial multiple-choice test was used to examine the negative testing effect. Thus, this way of examining the negative testing effect only dealt with participants in the same lure type condition. The means and standards deviations for this dependent variable can be found in Table 3.2.

Table 3.2. *Proportion of incorrect responses on the initial multiple-choice test that were given as responses on the final multiple-choice test for the same lure type condition as a function of passage status and multiple-choice question type.*

	Passage Status	
Initial MC Question Type	Read	Unread
A-D	.20 (.09)	.44 (.14)
A-E "E" Incorrect	.11 (.07)	.27 (.09)
A-E "E" Correct	.22 (.10)	.31 (.17)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

These proportions were subjected to a 2 (passage status: read, unread) X 3 (question type: A-D, A-E "E" Incorrect, A-E "E" Correct) repeated measures ANOVA. Participants responded with significantly more same lure responses to questions derived from unread passages (M = .34, SD = .13) compared to read passages (M = .18, SD = .09), $F(1,15) = 47.29$, MSE = .014. Also, there was a main effect of test type $F(2,30) = 10.80$, MSE = .013. Simple contrasts revealed that participants chose the same incorrect response on both tests more on A-D questions (M = .32, SD = .12) than on A-E "E" Correct questions (M = .27, SD = .14), $F(1,15) = 4.86$, MSE = .024. Furthermore, participants responded with the same incorrect response on both tests more for A-E "E" Correct questions than for A-E "E" Incorrect questions (M = .19, SD = .08), $F(1,15) = 4.72$, MSE = .33. These main effects were also qualified by a significant passage status by question type interaction, $F(2,30) = 5.13$, MSE = .009. The interaction showed that

participants responded with the same incorrect response on both tests to read questions of the A-D and A-E "E" Correct type, $F < 1$. However, participants provided the same incorrect response on both tests at a higher rate for A-D questions than A-E "E" Correct questions on questions derived from unread passages, $F(1,15) = 11.89$, $MSE = .025$.

Table 3.3. Proportion of correct responses on the initial multiple-choice test in Experiment 2 as a function of passage status and multiple-choice test question type.

	Passage Status	
MC Question Type	Read	Unread
A-D	.70 (.14)	.40 (.11)
A-E "E" Incorrect	.65 (.16)	.36 (.15)
A-E "E" Correct	.50 (.16)	.33 (.19)

Note. Standard deviations are provided in parentheses. MC = multiple-choice.

3.2.3 Initial Multiple-Choice Test Performance

As in Experiment 1, performance on the initial multiple-choice test was not a central interest. However, it was examined to see if the results obtained on the initial multiple-choice test in Experiment 1 would be replicated. A 2 (passage status: read, unread) X 3 (question type: A-D, A-E "E" Incorrect, A-E "E" Correct) repeated measures ANOVA was conducted on the proportion of correct responses on the initial multiple-choice test see Table 3.3. There was a main effect of question type, $F(1,31) = 14.53$, $MSE = .02$. Simple contrasts revealed that participants performed better on A-D

questions (M = .55, SD = .13) than on A-E "E" Incorrect questions (M = .51, SD = .16), although this was a marginally significant result, $F(1,31) = 3.84$, MSE = .027, p = .059. Furthermore, performance was significantly worse on A-E "E" Correct questions (M = .42, SD = .18) than on A-E "E" Incorrect questions, $F(1,31) = 10.01$, MSE = .052, and A-D questions, $F(1,31) = 27.94$, MSE = .039. Furthermore, there was a significant passage status by question type interaction, $F(1,62) = 6.15$, MSE = .013. Subsequent analyses revealed that participants performed better for questions taken from read passages on A-D questions opposed to A-E "E" Incorrect questions, $F(1,31) = 3.54$, MSE = .018. However, there was no difference in participants' abilities to correctly answer A-D and A-E "E" Incorrect questions derived from unread questions, $F(1,31) = 1.44$, MSE = .027, p = .24.

3.2.4 Summary of Experiment 2

The results from this experiment replicate and generalize the results obtained in Experiment 1. Performance on the final multiple-choice test demonstrated a positive testing effect for questions initially test with a correct answer as a response alternative (A-D and A-E "E" Incorrect questions). However, performance on final multiple-choice questions initially test in the A-E "E" Correct format did not show a positive testing effect. Surprisingly, performance on these questions was significantly lower than performance on questions that were not initially tested. This is a striking result that suggests that questions with a correct "none-of-the-above" response alternative not only negated the positive testing effect, but that they actually impair memory for initially tested material below that of material that was only studied. This result should be taken

with some caution since this was not significant in Experiment 1, and could be due to the increase in statistical power in the present experiment. However, the trend was similar in both experiments. Additionally, results of initial multiple-choice test performance replicated the results obtained in Experiment 1.

The results of the negative testing effect did not replicate Experiment 1. Participants in this experiment demonstrated a larger negative testing effect on A-D questions than A-E "E" Correct and A-E "E" Incorrect questions. This difference could possibly be due to the different operational definition afforded to the negative testing effect in each of the experiments. This difference is further discussed in the General Discussion.

4. GENERAL DISCUSSION

The purpose of the present experiments was the examine the influence of correct and incorrect "none-of-the-above response alternatives on the positive and negative testing effects. To do so, the paradigm used by Roediger and Marsh (2005) was adapted. In the preceding experiments, participants read passages, completed an initial multiple-choice test, and then took a final cued-recall (Experiment 1) or multiple-choice (Experiment 2) test. There were three types of questions on the initial multiple-choice test. A-D questions contained the correct answer and three incorrect lures and A-E "E" Incorrect questions contained these same alternatives plus an additional "none-of-the-above" response alternative. A-E "E" correct questions contained four incorrect alternatives and "none-of-the-above" as the most appropriate response alternative. The final test in both experiments contained the previously tested questions (critical questions) as well as questions that were about the passages but not tested on the initial multiple-choice test (control questions).

Although performance on the initial multiple-choice test was not the primary concern of the present experiments, the results concerning correct performance on the initial multiple-choice test replicated findings from previous research. First, questions with five response alternatives (A-E "E" Incorrect and A-E "E" Correct) were answered with fewer correct responses that questions with four alternatives. This finding is similar to Roediger and Marsh (2005) and Butler and colleagues (2006). However, in the present experiments, this detriment to performance was mostly confined to questions concerning studied material. Additionally, the results regarding performance on initial multiple-choice test questions with "none-of-the-above" as the most appropriate response alternative replicated the findings observed on target-absent lineups in the eyewitness identification literature (see Clark & Davey, 1995; Wells,

1993). In this literature, witnesses are more likely to choose a foil in a target-absent lineup than reject all of the suspects. In the present set of experiments, participants were very likely to select an incorrect A-D response alternative on questions with a correct "none-of-the-above" response alternative than reject all of the plausible alternatives and select "none-of-the-above". Moreover, these results demonstrate that the same inclusive response alternative (i.e., "none-of-the-above") can lead to differential levels of performance on an initial test.

4.1 The Positive Testing Effect

The major focus of these two experiments was to examine the positive testing effect with an interpolated multiple-choice test that included questions with a correct and incorrect "none-of-the-above" response alternative. It was predicted that a positive testing effect would be observed on questions that contained the correct answer, regardless of the presence of a "none-of-the-above" response alternative. This hypothesis was supported by the present results. Participants in both experiments provided more correct answers to questions that were initially tested in the A-D or A-E "E" Incorrect format than to questions that were not initially tested. This finding replicates other previous studies (e.g., Butler et al., 2006; Roediger & Marsh, 2005; Whitten & Leonard, 1980).

Additionally, it was further predicted that initial multiple-choice test questions with "none-of-the-above" as the most appropriate response would not demonstrate a positive testing effect. In Experiment 1, participants performed statistically similar on cued-recall questions initially tested in the A-E "E" Correct question format and control questions. In Experiment 2, this difference became statistically different. Participants in the second experiment provided significantly fewer correct answers to final multiple-choice test questions that were initially tested in the A-E "E" Correct format than

control questions. Importantly, this occurred even when novel incorrect lures were included on the final multiple-choice test in place of the incorrect lures used on the initial multiple-choice test. These findings are important because they demonstrate that the type of questions given on an initial multiple-choice test can moderate the testing effect. As with initial multiple-choice test performance, these results provide evidence that questions on an initial test with a "none-of-the-above" alternative can have differential effects on the positive testing effect depending on the correctness of the "none-of-the-above" alternative. Theoretical implications for these results will be discussed later.

This finding fits well with the argument posed by Schooler and colleagues (1988). As stated earlier, these authors argued that tests that promote incorrect responding have deleterious effects on memory. These authors demonstrated that forcing participants to respond incorrectly can block access to the correct information. Other researchers examining the misinformation effect have supported this notion by showing the importance of commitment to misinformation in producing the misinformation effect (Schreiber & Sergent, 1998; Zaragoza, Payment, Ackil, Drivdahl, & Beck, 2001). Interestingly, the positive testing effect was not present even when a multiple-choice test was given as the final retention test and the incorrect lures were different from the initial test. This finding replicated and extended the results from Schooler and colleagues (1988, Experiment 2). Taken together, the findings from both experiments provide conclusive evidence that multiple-choice questions with a correct "none-of-the-above" alternative will not lead to a positive testing effect.

4.2 The Negative Testing Effect

The design of these experiments allowed for the examination of the negative testing effect. Indeed, participants in both experiments did import incorrect lures on the

initial multiple-choice test as responses to both final cued-recall and multiple-choice test questions. This finding replicated and extended previous research (Butler et al., 2006; Roedger & Marsh, 2005; Whitten & Leonard, 1980). It was hypothesized that questions with a correct "none-of-the-above" response alternative would demonstrate the largest negative testing effect. The results from Experiment 1 supported this hypothesis. Participants responded with more incorrect lures on final cued-recall questions initially tested with a correct "none-of-the-above" alternative relative to questions with an incorrect "none-of-the-above" alternative and questions without this alternative. However, the trend in the negative testing effect observed in Expderiment 1 did not generalize to the new measure used to examine the negative testing effect in Experiment 2. Participants in Experiment 2 showed a larger negative testing effect on A-D questions than the other two question types with a "none-of-the-above" alternative.

The different results obtained between these two experiments should be taken lightly. First, the measures used to examine the negative testing effect were different for both experiments. In Experiment 1, the negative testing effect was operationally defined as the proportion of incorrect multiple-choice lures that appeared as responses on final cued-recall test questions. In Experiment 2, the negative testing effect was measured by examining the proportion of incorrect responses on the final multiple-choice test that were the same as the initial incorrect responses on the initial test. These two measures are fundamentally different in the sense that one is dependent upon only seeing the incorrect lures on the initial multiple-choice test and importing them onto the final test (Experiment 1) and the other is dependent upon selecting the same incorrect lure on both the initial and final tests (Experiment 2). One way to equate these two findings would be to change the measure used in Experiment 2. The other measure that could have been used in this experiment was the proportion of questions that were incorrectly

answered for participants in the same lure condition. This measure would be similar to the one used in Experiment 1 since it includes exposure to the incorrect lures, not just commitment. Using this new measure, it is obvious that the negative testing effect was increased in Experiment 2.

4.3 Theoretical Implications: The Retrieval Hypothesis

The present results can be explained by the accessibility and retrieval blocking hypothesis (Darley & Murdock, 1971; Raaijmakers & Shiffrin, 1981; see however Belli, 1993). In the misinformation literature, this hypothesis states that the initial retrieval of misinformation can block access to the correct information. As discussed earlier, an initial retrieval episode (i.e., a multiple-choice test) can increase the accessibility of studied material by blocking access to competing information. For instance, participants in the present experiments showed a positive testing effect on items that had the correct response since they were more likely to choose the correct answer. The higher level of performance on A-D and A-E "E" Incorrect questions on the initial multiple-choice test supports this theoretical standpoint (see Tables 4 and 7; see Figures 1 and 2). On these questions, participants were very likely to retrieve the correct answer to the question. Recognition of the correct answer could have been due to either recollection of the correct answer or the level of familiarity associated with the correct answer. In regards to the accessibility and retrieval blocking hypothesis, the retrieval of the correct answer led to increasing accessibility of that information and mental blocking of the non-retrieved lures. This finding is similar to Chan and McDermott (in press) who demonstrated that testing increases estimates of recollection.

The above notion can help explain why a positive testing effect was not observed on questions initially tested with a correct "none-of-the-above" alternative. As discussed above, recollection and familiarity can act in concert to direct a person in

selecting the correct answer. However, this is not the case when "none-of-the-above" is the correct response alternative. A participant may have chosen this response because the four incorrect lures had of a level of familiarity lower than his or her response criterion, and thus were not given as answers. Additionally, a participant may have chosen "none-of-the-above" because they recalled the correct answer that was not one of the answer choices. However, the recalled answer may have been incorrect in regards to the question (Gross, 1994). For example, lets say a person is posed with the question *"What was Louis Armstrong's Nickname?"* where "none-of-the-above" is the most appropriate response. Of course, the person may accurately recall that his nickname was *"Satchmo"* and correctly select "none-of-the-above" this way. However, this person may incorrectly recollect that his nickname was *"Lou"* and select "none-of-the-above" because "*Lou*" was not a response alternative. Even though this person may have correctly accepted "none-of-the-above', they committed themselves to another incorrect answer. This type of incorrect responding supports the retrieval blocking hypothesis in that recalling an incorrect answer to a question can still block access to the correct information.

The present results do not support the elaborative retrieval hypothesis. This hypothesis would predict a larger positive testing effect for questions initially tested with a correct "none-of-the-above" response alternative. As noted above, these types of questions require accurate recall of the answer, which is more difficult than just recognizing the answer. Even though this result was not obtained, the present paradigm provides a useful tool to examine this effect and will be discussed later.

4.4 Limitations and Future Directions

There are some limitations to the present experiment. Most notably is the lack of a delay condition. Many teachers do not give tests back to back. Instead, most teachers

give tests weeks or months apart. As many researchers have noted (e.g., Spitzer, 1939; Roediger & Karpicke, 2006a, 2006b) the positive testing effect is most pronounced over a delay. A delay condition may have demonstrated a positive testing effect for questions with "none-of-the-above" as the most appropriate response alternative. However, this is most likely not the case. Zaragoza and colleagues (2001) demonstrated that initial forced confabulation to misinformation showed a large misinformation effect after a month's delay. Their findings would lead to the prediction that the positive testing effect would be negated after a delay for questions initially tested with a correct "none-of-the-above" alternative. Future research that examines "none-of-the-above" and the testing effects should include a delay condition.

Another possible limitation of the present study was that it lacked a restudy control condition. However, the lack of including this condition does not minimize the results obtained in both experiments. The design of both experiments used a within-participant manipulation for a control condition. Additionally, both experiments used an immediate retention test to examine the positive and negative testing effects. The inclusion of a restudy condition would have confounded the present results since restudying the non-tested material would have most likely boosted performance on the final test above the tested items. This would be predicted based on Karpicke and Roediger's (2007) results. They demonstrated that receiving more study sessions boosted performance on the final retention test given five minutes the initial study/test session. Thus, the inclusion of this condition would have masked the positive testing effect and confounded the results obtained in the present experiments.

As noted earlier, the present paradigm provides a useful tool to examine the elaborative retrieval explanation of the testing effect. As pointed out by Carpenter and DeLosh (2006), cues that require a more extensive search of memory led to a greater

positive testing effect. Questions that contained a correct "none-of-the-above" response alternative should, and most likely do, require more elaborative retrieval processes to be accurately recollected. However, it does not seem like participants in the present experiments used an elaborative retrieval process. Instead, it appears that they relied on the familiarity, or the lack there of, of the response alternatives.

Odegard and Lampinen (2005) discussed that students should utilized a recall-to-reject strategy while taking class examinations since it would benefit subsequent memory performance. It would be interesting to see if participants would show a positive testing effect when they are directed to use this type of strategy. One way this could be done is by having participants provide the factually correct answer when they provide "none-of-the-above" as a response. This may aid the participants by facilitating recall of the correct answer. Additionally, if this were the case, these items would have a larger generative component and should show a normal, if not larger, positive testing effect.

4.5 Pedagogic Implications

The present results have implications for educational instruction. The present research adds to the growing literature that suggest that tests should be utilized more often in class instruction since they promote learning and can strengthen memory (Bangert-Drowns et al., 1991; Dempster 1992; 1997; Dempster & Perkins, 1993; Roediger & Karpicke, 2006b). However, the results obtained in both of the experiments suggest that caution should be taken in when constructing multiple-choice test questions. Educational research regarding the use of "none-of-the-above" as a response alternative has been sparse with many researchers depating whether or not these types of questions should be used (Gross, 1994; Knowles & Welch, 1992). Gross (1994) argued that questions with a correct "none-of-the-above" response alternative minimize

the chances of recalling the correct answer and suggested that these types of questions are flawed with negative consequences. The results of both experiments are in line with this suggestion. Based on the present results, correct "none-of-the-above" response alternatives should not be utilized in class instruction not because they are more difficult to answer (Knowles & Welch, 1992), but because they can impair memory and subsequent test performance.

4.6 Conclusion

The present experiments demonstrated that different types of multiple-choice test questions have differential effects on both the positive and negative testing effect. Initial multiple-choice questions with a correct "none-of-the-above" response alternative did not show a positive testing effect on a final retention test. These types of questions also demonstrated a larger negative testing effect on the final retention test. Many researchers have pointed out that the positive testing effect outweighs the negative testing effect (e.g., Marsh et al., 2007; Roediger & Karpicke, 2006a; Roediger & Marsh, 2005). As discussed by Schooler and colleagues (1988), tests that promote incorrect responding have detrimental effects on memory for the tested material on later tests (see also Schreiber & Sergent, 1998; Zaragoza et al., 2001). Indeed, participants in the present experiments were poor at selecting the correct response on the initial multiple-choice test. Additionally, participants in these experiments were also poor and providing the correct answer on final retention test questions when they were initially tested with a correct "none-of-the-above" alternative. These results do not support the use of "none-of-the-above" as alternatives on multiple-choice test. As demonstrated by both of the experiments reported here, multiple-choice questions with a correct "none-of-the-above" response alternative yielded large memorial consequences for the tested material because participants were biased to select an incorrect alternative.

REFERENCES

Armitage, S.G. (1945). An analysis of certain psychological tests used for the evaluation of brain injury. *Psychological Monographs*, *60*, 1-48.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89-99.

Belli, R. F. (1993). Failure of interpolated tests in inducing memory impairment with final modified tests: Evidence unfavorable to the blocking hypothesis. *American Journal of Psychology, 106,* 407-427.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.) *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Brainerd, C. J., Kingma, J., & Howe, M. L. (1985). On the development of forgetting. *Child Development, 56,* 1103-1119.

Brown, A. S. (1988). Encountering misspellings and spelling performance: Why wrong isn't right. *Journal of Educational Psychology, 80,* 488-494.

Brown, A. S., Schilling, H. E. H., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology, 91,* 756-764.

Bruning, R. H. (1968). Effects of review and testlike events within the learning of prose material. *Journal of Educational Psychology, 59,* 16-19.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941-956.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514-527.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619-636.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268-276.

Chan, J. C. K., & McDermott, K. B. (in press). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553-571.

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior, 29,* 151-172.

Cull, E. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215-235.

Darley, C. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91,* 66-73.

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30,* 287-307.

Dempster, F. N. (1992). Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education, 25,* 213-217.

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.) *Handbook on testing* (pp. 332-346). Westport, CT: Greenwood Press.

Dempster, F. N., & Perkins, P. G. (1993). Revitalizing classroom assessment: Using tests to promote learning. *Journal of Instructional Psychology, 20,* 197-203.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology, 6,* 217-226.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75,* 309-313.

Dysart, J. E., Lindsay, R. C. L., Hammond, R., & Dupuis, P. (2001). Mug shot exposure prior to lineup identification: Interference, transference, and commitment effects. *Journal of Applied Psychology, 86,* 1280-1284.

E-Prime (Version 2005.1.1.4.1) [Computer program]. Pittsburg, PA (http://www.pstnet.com): Psychology Software Tools.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80,* 179-183.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213-216.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392-399.

Gorenstein, G. W., & Ellsworth, P. C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology, 65,* 616-622.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of "none-of-the-above". *Evaluation & The Health Professions, 17,* 123-126.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30,* 67-80.

Higham, P. A., & Gerrard C. (2005). Not all errors are created equal: Metacognition and changing answers on a multiple-choice test. *Canadian Journal of Experimental Psychology, 59,* 28-34.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562-567.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649-667.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading correctly and incorrectly spelled words. *Canadian Journal of Psychology, 44,* 345-358.

Jones, H. E. (1923-1924). The effects of examination on the performance of learning. *Archives of Psychology, 10*, 1-70.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528-558.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151-162.

Knowles, S. L., & Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above". *Educational and Psychological Measurement, 52,* 571-577.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology, 67,* 259-266.

Loftus, E.F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*, 361-366.

Loftus, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19-31.

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory, 7,* 79-90.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194-199.

McDaniel, M. A. & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371-385.

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34,* 261-267.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 596-606.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74,* 18-22.

Odegard, T. N., & Lampinen, J. M. (2005). Recollection rejection: Gist cueing of verbatim memory. *Memory & Cognition, 33,* 1422-1430.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning, 4,* 11-18.

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect of true-false examination questions. *Journal of Educational Psychology, 17,* 52-56.

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 14,* 249-255.

Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181-210.

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155-1159.

Runquist, W. (1986a). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology, 40,* 65-76.

Runquist, W. (1986b). Changes in the rate of forgetting produced by recall tests. *Canadian Journal of Psychology, 40,* 282-289.

Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition, 16,* 243-251.

Schreiber, T. A., & Sergent, S. D. (1998). The role of commitment in producing misinformation effects in eyewitness memory. *Psychonomic Bulletin & Review, 5,* 443-448.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592-604.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 716-727.

Spitzer, H. F. (1939). Studies in retention. *Journal of Education Psychology, 30,* 641-656.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210-221.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true-false examinations. *Journal of Educational Research, 83,* 119-124.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research, 86,* 357-362.

Tulving, E. (1967). The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175-184.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381-391.

Underwood, B. J. (1970). A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 9,* 573-580.

Wells, G. W. (1993). What do we know about eyewitness identification? *American Psychologist, 48,* 553-571.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571-580.

Whitten, W. B., II, & Bjork, R. A. (1977). Learning from tests: effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465-478.

Whitten, W. B., II, & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 127-134.

Zaragoza, M. S., Payment, K. E., Ackil, J. K., Drivdahl, S. B., & Beck, M. (2001). Interviewing witnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science, 12,* 473-477.

BIOGRAPHICAL INFORMATION

Joshua Koen has been a student at the University of Texas at Arlington since 2003. His aspiration is to obtain a doctoral degree in experimental psychology and become a professor at a university. This research project culminated in a publication in *Memory*. Joshua Koen is involved in other research projects that are examining the role of plausibility in editing memory in young adults, the role of plausibility in memory editing in children, and the neural underpinnings of recall processes in memory. His research interests include theoretical and applied research on human learning and memory, cognitive function in normal and abnormal aging populations, the effects of emotionally arousing stimuli on memory, dual-process theories of recognition, the testing effect, and cognitive neuroscience.